ARTICLE IN PRESS JAL:444

Journal of Applied Logic ••• (••••) •••–•••

Contents lists available at ScienceDirect

Journal of Applied Logic

www.elsevier.com/locate/jal

Hybrid evolutionary workflow scheduling algorithm for dynamic heterogeneous distributed computational environment

D. Nasonov [a], A. Visheratin [a], N. Butakov [a], N. Shindyapina [a], M. Melnik [a], A. Boukhanovsky [b]

[a] *ITMO University, Saint-Petersburg, Russian Federation*
[b] *Netherlands Institute for Advanced Study in the Humanities and Social Sciences, Wassenaar, The Netherlands*

A R T I C L E I N F O

Article history:
Available online xxxx

Keywords:
Workflow scheduling
Dynamic environment
Evolutionary algorithms
LEFT
GA
GAHEFT

A B S T R A C T

The optimal workflow scheduling is one of the most important issues in heterogeneous distributed computational environments. Existing heuristic and evolutionary scheduling algorithms have their advantages and disadvantages. In this work we propose a hybrid algorithm based on heuristic methods and genetic algorithm (GA) that combines best characteristics of both approaches. We propose heuristic algorithm called Linewise Earliest Finish Time (LEFT) as an alternative for HEFT in initial population generation for GA. We also experimentally show efficiency of described hybrid schemas GAHEFT, GALEFT, GACH for traditional workflow scheduling as well as for variable workload in dynamically changing heterogeneous computational environment.

1. Introduction

Nowadays scientific progress cannot be imagined without the usage of high performance distributed computational environments, such as clusters, grids, private and public clouds. Usually workflow formalism is used for solving complex scientific tasks in such environments within different kinds of workflow management systems (WMS) [1,19,20]. One of the most important functional aspects of WMS operation is workflow scheduling. Efficient scheduling within a distributed computational environment allows time, cost, energy and other critical factors to be optimized. Proper tasks scheduling can be crucial for urgent computation systems, such as flood early warning systems [18]. Determination of an optimal workflow plan is an NP-complete problem. This fact along with heterogeneity of computational resources, diversity of computational models, different constrains defined by a user, and constantly increasing complexity of the environment empowers many groups of researchers to propose new solutions in this domain area.

E-mail addresses: denis.nasonov@gmail.com (D. Nasonov), alexvish91@gmail.com (A. Visheratin), natavts@gmail.com (N. Butakov), avb_mail@mail.ru (N. Shindyapina).

http://dx.doi.org/10.1016/j.jal.2016.11.013
1570-8683/

Traditionally, a number of heuristic approaches in workflow scheduling are used. The most common of them are list-based algorithms such as CPOP, HCPT, HPS, PETS [2] and widely applied heuristic – Heterogeneous Earliest Finish Time (HEFT) [3], which also performs better than others greedy heuristics for most cases [2]. All of them are quite fast and provide suitable solutions; however, they are only locally optimal. For example, HEFT runtime for Montage workflow with 100 tasks (Montage_100, see Section 4 for details) in our experiments is about 30 ms, which does not much influence workflow execution time.

Meta-heuristic algorithms compose another class of techniques, which are used for workflow scheduling. Among them: GRASP [4], genetic algorithm (GA), memetic algorithm (MA) [5], ant colony optimization (ACO), particle swarm optimization (PSO) and simulated annealing (SA) (a survey of using these algorithms can be found in [6]). GA is the most frequently applied in the context of workflow scheduling and better adapted for combinatorial optimization than PSO, ACO and is able to generate more sophisticated solutions than simpler GRASP and SA. It allows a high-quality solution to be derived from a large search space by applying the evolutionary principles [7]. Potentially, all meta-heuristics can provide better solutions than traditional heuristics in terms of constraints satisfaction, but they dramatically lose in computational time. In our experiments, for example, it took about 18 s for GA with population size of 100 and 300 iterations to complete in case of Montage_100, which is quite notable time for workflow with makespan of about 500 s. The growing interest of the science community [15] in nature-inspired, and in particular, evolutionary meta-heuristics algorithms and their recent success in practical applications (see [16,5] for examples), motivate searching for new ways to use them to improve performance of schedulers in distributed environments.

In this paper we describe a new hybrid algorithm that combines the best characteristics of heuristic algorithms (HEFT and proposed LEFT) and the meta-heuristic algorithm (GA) for scheduling workflows in a dynamically changing distributed computational environment. The aim of the hybrid approach is to improve quality of solutions in comparison with heuristics ones making it closer to the optimum while saving the same level of consumed computational time spent on the generation.

2. Background

Usually a scientific workflow is represented as a Directed Acyclic Graph (DAG), where nodes are computational tasks and edges determine data dependencies (data transfers) or control dependencies (order of execution). The detailed definition of a DAG is described in [8]. The objective for workflow scheduling procedure is usually defined in the form of a makespan minimization [9]. However, in real environments, there can be other constraints and objectives such as resources usage cost or more complex cases where several constraints are taken into account [17]. Taking into account data transfer overheads is critical in case of data-intensive computations. For example, the use of one less powerful resource for two connected tasks (no transfer overheads) would be more efficient than the use of two more powerful separated resources due to significant data transfer time. Distributed computational environment consist of heterogeneous resources with differences in performance, network bandwidth, access policies, etc. Moreover, resources can crash during the computational process and restored over time. Also, due to a stochastic nature of the distributed environment, there is no reliable way to predict the execution time of a task precisely.

2.1. Workflow scheduling algorithms

It has already been mentioned that heuristic and meta-heuristic algorithms have their pros and cons. The genetic algorithm has crucial limits, which prevent its universal usage for tasks scheduling. These limits are: (a) a fitness function with expensive evaluation; (b) no guarantees that the optimal solution can be found in a limited time; (c) the GA's solution can be less efficient than the solution provided by the heuristic algorithm; (d) the solution provided by the GA can be locally optimal; (e) the GA's optimal parameters